Duplicate Document Detection in a Web Crawler System

ABSTRACT OF THE INVENTION

Duplicate documents are detected in a web crawler system. Upon receiving a newly crawled document, a set of documents, if any, sharing the same content as the newly crawled document is identified. Information identifying the newly crawled document and the selected set of documents is merged into information identifying a new set of documents. Duplicate documents are included and excluded from the new set of documents based on a query independent metric for each such document. A single representative document for the new set of documents is identified in accordance with a set of predefined conditions.